SYNOPSIS

PhD. in Zoology

Identification of novel genes, pathways and mutations associated with Acute Myeloid Leukemia

Submitted by

R.Nithya

(19PHZOF001)

Supervisor

Dr. K.S. Santhy

Professor

Department of Zoology

Submitted to

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore–641043

September, 2022

INTRODUCTION

Acute Myeloid Leukemia (AML) is a common type of hematologic neoplasm characterized by molecular characteristics and its incidence rate is increasing every year (Estey, 2020). Ameriacn Cancer Society in 2019 recorded 19,950 new cases of AML and 10,430 deaths due to AML in USA (Siegel *et al*., 2021). The estimated overall survival of AML is 8.5 months; the 2-year and 5-year survival is 32% and 24% (Vakiti *et al*., 2021). Indeed, genetic factors play an important role in cause of leukemia. On the other hand, smoking, ionizing radiation, prior chemotherapy, Down's syndrome and exposure to chemicals such as benzene are major risk factors for leukemia (Tebbi, 2020).

An earlier study suggested that mutations in the molecular pathogenesis of AML could be divided into five categories. Class I mutations, such as FLT3, RAS, JAK2, and KIT, are associated with proliferation and survival. Class II mutations are associated with blocked differentiation, such as RUNX1 – RUNX1T1, CBFB – SMMHC, CEBPA, GATA1, and GATA2. Class III mutations including IDH1/2, DNMT3A, TET2, ASXL1, EZH2 and UTX are linked with epigenetic modifications. Class IV mutations are related to adhesion and cell – cell interaction. Class V mutations are associated with DNArepair/RNA-splicing (Im *et al*., 2014). Indeed, molecular analysis of leukemic blasts from AML patients suggested that there was an obvious heterogeneity in the gene expression. However, the detailed pathophysiology of the AML still remains unclear, especially the complicated molecular mechanisms (Handschuh *et al*., 2019). Therefore, identification of more effective biomarkers and biological pathways, and better understanding of the mechanism of AML is needed.

Identifying cancer-promoting genes, pathways, and driver mutations promises to understand better the mechanisms underlying the disease and provide potential diagnostic and therapeutic markers. However, experimental characterization of all identified mutations is not practical and is usually time-consuming, costly, and labor-intensive. And it is difficult to distinguish disease-associated synonymous mutations from benign ones. Several excellent bioinformatic platforms, algorithms and tools have been proposed to support the prioritization of novel genes, pathways and disease prone sites to meet this need.

Due to revolution in gene sequencing technology and emergence of open data-sharing platforms such as GEO, TCGA and COSMIC database, an increasing number of ongoing studies

are aimed at defining the gene expression profile associated with cancers. Moreover, the molecular aspect of malignancy is complicated, which is connected with the deletion and modification of tumor suppressor genes and the provocation of oncogenes.

In this context, the present study used a series of bioinformatic methods and identified DEGs, GO, pathway enrichments, PPI network and hub genes associated with PTPN11 mutation. Then, we considered a weighted gene co-expression network, identified the key gene modules that affect the pathogenesis of AML, conducted enrichment analysis, and screened out six new biomarkers. Nevertheless, identifying the potential residues associated with AML is a key to the early detection of cancer pathogenesis, and it helps to develop personalized therapeutics. Emerging sequencing technologies have enhanced the low mutational burden in myeloid leukemia, and these recurrent mutations have been recognized as drivers. However, in recent years much work has pointed out the critical role of driver mutations in human diseases. Though there are few computational methods such as VEPAD and CanProSite have been proposed to predict disease/cancer-prone sites for interpreting confirmatory principles of driver mutations in Alzheimer's and Lung Cancer (Rangaswamy et al., 2020, Pandey & Gromiha, 2021), there is no such methods developed for AML prediction using driver mutation. Hence, an accurate prediction tool is essential to identify the AML cancer-prone sites. So we classified the disease-causing mutations and developed a computational model to predict the driver mutations in AML.

Several bioinformatic studies had been reported on AML. But those studies has several limitations:

- Those studies are based on only the sample groups (disease and normal) and the results are not applied to understand the impact of specific gene mutation on pathways and gene expression.
- Previous studies reported the co-expressed genes associated with karyotype, blood cell counts and mutation and not attempted to define the co-expressed gene patterns associated with blast cell percentage.
- In addition, there is no specific method available for AML for identifying the disease prone sites with high accuracy.

Considering all the above information, the present investigation was undertaken with the following objectives.

- To identify the impact of PTPN11 mutation on differential gene expression and prognosis of AML patients.
- To identify co-expressed genes that are highly correlated with blast cell percentage and overall survival.
- To classify potential disease-causing mutations in AML using machine learning algorithms.
- To statistically evaluate disease-causing and neutral mutations in AML.
- To develop a web server for predicting the disease-prone sites of AML.

MATERIALS AND METHODS

The methods and tools used in the study are as follows:

I. Identification of Differentially Expressed Genes and Key Pathways Associated with PTPN11 Mutation

A. Data collection

To compare genes and mRNA expression between PTPN11 mutation and wild-type adult AML patients, mRNA expression pattern and corresponding survival profiles were obtained from The Cancer Genome Atlas (TCGA) database

B. Identification of DEGs

EdgeR (Robinson, 2010) was used to screen DEGs between PTPN11 mutation and wild-type AML samples according to the user's guide. DEGs were identified with the cut-off value of $\log2|\text{fold change (FC)}| \geq 1$ and P-Value < 0.05. A heat map and volcano plot of DEGs were drawn by the ggplot2 package in the R platform.

C. Functional annotation and pathway enrichment analysis of DEGs

Database for Annotation, Visualization and Integrated Discovery (DAVID) was used to annotate input genes, classify gene functions, identify gene conversions, and carry out Gene

Ontology (GO) term analysis (Sherman *et al*., 2022). To identify the DEG's functional annotation, we analyzed GO terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment with DAVID, by specifying a P-value < 0.05 for statistical significance.

D. Gene Set Enrichment Analysis

To investigate the effect of PTPN11 mutations on various biological function gene sets in adult AML patients, differences in gene mRNA expression levels of biological functional annotation and pathways between PTPN11 mutation and wildtype patients were analyzed by GSEA v2-2.2.3. Reference gene sets from the Molecular Signatures Database (MSigDB) of c2 (c2.cp.kegg.v5.2.symbols.gmt) and c5 (c5.bp.v5.2.symbols.gmt; c5.mf.v5.2.symbols.gmt; c5.cc.v5.2.symbols.gmt; consist of genes annotated by the same GO terms). The MSigDB of c2 is a pathway gene set, which was curated from publications and extracted from canonical pathways and experimental signatures, whereas the MSigDB of c5 was constructed on genes annotated by the same GO terms. The number of permutations was set at 1,000. Enrichment results satisfying a nominal P-value cutoff of 0.05 with a false discovery rate (FDR),0.25 were considered statistically significant.

E. Integration of PPI network

The Search Tool for the Retrieval of Interacting Genes (STRING) database is an online tool to evaluate protein-protein interaction information. To evaluate the Protein-Protein interaction relationships among DEGs, we mapped the DEGs to STRING to evaluate the validated interactive relationships among DEGs. Experimentally validated interactions with a combined score>0.4 were selected as significant. Using the PPI networks. The cytoHubba plugin in Cytoscape (Shannon, 2003) was performed to identify hub genes of the PPI network with defaults. GO enrichment terms of hub genes and genes in modules were also analyzed by Metascape (Zhou *et al*., 2019).

F. Statistical analysis

All the statistical analyses were conducted with Graph Pad Prism version 9.0. The t-test was used to evaluate the gene expression level between PTPN11 mutation and wild-type AML. The Kaplan-Meier was used to calculate the overall survival of patients Hazard Ratio. P-value < 0.05 was indicated as statistically significant.

II. Identification of co-expression networks, key genes and pathways in Acute Myeloid Leukemia

A. Data collection and pre-processing

Gene expression data of AML samples were collected from the microarray database GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc).

B. Construction of gene co-expression network and hub-gene selection

Construction of gene co-expression network was performed by using R software with WGCNA package. β was a soft threshold function that can emphasize a strong co-relation and make the network conform to the standard scale-free network. The function provides an appropriate β value ranging from 1 to 20. Then, through hierarchical clustering, co-expression modules were identified, and the tree was obtained. The minimum number of module genes was set as 30, and the modules were merged for the second time accordingly to the modules with a correlation more significant than 75%.

C. Functional enrichment analysis in the key modules

Functional enrichment analysis was performed on genes in the significant modules using the DAVID (https://david.ncifcrf.gov/tools.jsp) web server (Jiao et al., 2012). Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathway analysis was employed for annotating genes and gene products for identifying novel biological processes. The p-value < 0.05 was used as the cut-off criterion.

D. Identification of hub genes

The correlation weight value from each gene in the dark red and magenta modules was generated with the WGCNA package (Langfelder & Horvath, 2008). The gene correlation was visualized using Cytoscape software by the cyto-Hubba plugin. The hub genes that were highly interconnected with nodes in a module were identified with the degree of connectivity method. The top 10 genes with a correlation weight value > 0.1 were defined as hub genes.

E. Validation of hub genes

GEPIA (http://gepia.cancer-pku.cn/) database was used to validate the expression difference of candidate genes among AML samples (Tang *et al*., 2017). Kaplan-Meier survival plots were applied to validate the hub genes prognosis.

III. Analysis and identification of disease prone sites in Acute Myeloid Leukemia using machine learning algorithms

The mutation data set for AML was constructed from cosmic database and discriminated the disease-causing and neutral mutation using an ensemble of algorithms. We have analyzed a set of 735 cancer-prone and 1328 neutral sites in AML using sequence-based features, namely amino acid properties, PSSM profiles, motifs, secondary structure, conservation, and topological location of the mutation.

Multiple machine learning algorithms were utilized to prioritize cancer-prone and neutral mutations.

IV. Statistical analysis of Disease-causing and Neutral mutations in Acute Myeloid Leukemia

Frequency of amino acids, di-and tri-peptide motifs in disease-causing and neutral sites in AML were observed. Further, we grouped the mutations based on different topological regions of protein namely outside region, inside, membrane, and signal peptide sequence regions.

V. Web application development

Front end web interface is developed using react JS which is very light weight and fast as it falls under single page application. The backend is developed using Python Flask combination and the database is built on SQLite.

FINDINGS OF THE STUDY

I. Identification of Differentially Expressed Genes and Key Pathways Associated with PTPN11 Mutation

- A total of 3268 DEGs (1427 upregulated and 1841 downregulated) were identified.
- Biological processes, DEGs suggested significant enrichment in apoptotic processes. For cellular components and molecular function, DEGs were enriched in the cytoplasm and protein binding.
- KEGG pathway analysis showed that DEGs were highly enriched in metabolic pathways.
- GSEA analysis of GO enrichment showed that biological processes such as regulation of telomeres via telomere lighting vesicle budding from membrane, DNA biosynthetic

process, positive regulation of DNA metabolism process, telomer organization were significantly enriched.

- ➢ The GO enrichment analysis of molecular function was significantly enriched in ubiquitin like protein binding, cadherin binding, mRNA binding, cell adhesion molecular binding. Furthermore, the cellular component was enriched for inner mitochondrial membrane protein, procatalytic splisome, ribonucleic complex, and mitochondrial protein complex.
- ➢ In the GSEA analysis of KEGG pathways, the PTPN11 mutation group was associated with the antigen processing and presentation, MAPK-signalling pathway, Insulin signalling pathway, Ubiguitin mediated proteolysis.
- ➢ The PPI network analysis identified some PTPN11 mutation-associated hub genes. The top 10 genes ranked by degree of connectivity were identified as hub genes, including GAPDH, HSP90AA, MYC, UBB, HSPA4, EP300, CCDS, POLR2A, EFTUD2, RPL11. GAPDH had the highest degree of nodes.
- ➢ GO analysis of Metascape suggests that hub genes are significantly enriched in metabolic, developmental, and biological processes which have been associated with cancer.
- ➢ Overall survival analysis for AML patients grouped by PTPN11 mutation revealed that AML patients with PTPN11 mutation had a shorter median survival time (MST) than those without PTPN11 mutations.
- ➢ Multivariate Cox proportional hazards regression analysis suggested that PTPN11 mutations were significantly associated with a poor clinical outcome and an increased risk of death, compared to these patients without PTPN11 mutations

II. Identification of co-expression networks, key genes and pathways in Acute Myeloid Leukemia

- ➢ By bioinformatic method WGCNA, the genes were clustered into seven modules, of which two modules were the most closely related to the development and progression of AML. The scale-free network condition was satisfied when $\beta = 14$, the scale independence value achieved was 0.14 and which reveals lower connectivity.

- ➢ Hierarchical clustering dendrogram grouped the genes with similar expressions into the same module. A total of 1009 and 493 mutual gene probes were clustered into dark red and magenta gene modules.
- ➢ GO analysis of all the genes in these two modules showed that the apoptotic process, positive and negative regulation of transcription, signal transduction, and protein phosphorylation was the most abundantly represented in the biological process category.
- ➢ In the cellular component category, cytosol, nucleus, cytoplasm were the top three terms, while protein binding, metal ion binding and ATP binding activity were top molecular function terms.
- ➢ KEGG enrichment analysis revealed that co-expressed genes were enriched in MAPK signaling pathway (hsa04010) and Pathways in cancer (hsa05200).
- ➢ Ten genes namely AKT1, MYC, JUN, ATM, NOTCH1, PTPN11, ITGB1, MAPK14, TP53BP1, NFKBIA were identified as hub genes.
- ➢ Kaplan Meier plots indicated that ATM, MYC, NOTCH1, PTPN11 genes were upregulated whereas ITGB1, JUN genes were downregulated and found to be correlated with poor overall survival in AML patients.

III. Analysis and identification of disease prone sites in Acute Myeloid Leukemia using machine learning algorithms

- ➢ In feature selection method we eliminated the least important features and our final model utilized single feature tri-peptide motif (*M*).
- ➢ The study identified that all the methods discriminate the AML cancer and neutral sites with an accuracy range of 80% and the AUC lies at 93%. We have examined the performance of other machine learning techniques, Xgboost, Adaboost and Neural networks, and we observed that the Xgb model outperforms other classifiers with the accuracy of 80 % and AUC of 0.89. The Other machine learning models attained the maximum accuracy and AUC of 77 % and 0.81, respectively

IV. Statistical analysis of disease prone sites and neutral sites in Acute Myeloid Leukemia.

- The present study observed that Gly and Cys are dominant in cancer prone sites, whereas Ile and Val are preferred in neutral sites.
- Specifically the di-peptide (nM: n is any residue and M is mutant residue) AG is the most preferred motif, 3 at cancer prone sites and 4 at neutral site followed by GG,GR,SA,SR,NR,TR,RD,GC and VR.
- On the other hand, the motif Mn is preferred by the dipeptides RL followed by GV,GG,RN,RR,CA,PP,RS,GL and RP.
- Although Arg is not a preferred residue in myeloid cancer prone sites, its combination with Gly(GR) is predominantly present with an odds ratio of 2.73. On the other hand, the motif Mn is preferred by the dipeptides RL with an odds ratio of 6.44 followed by GV,GG,RN,RR,CA,PP,RS,GL and RP. Intestingly, Gly is the most preferred residue at myeloid cancer prone sites and several dipeptide motifs are residues in combination with Gly.
- Considering the motif n_M(n: any residue; gap and M: mutant residue), L_R is present in myeloid cancer prone sites with an odd ratio of 3.26. Other preferred specific motifs in myeloid cancer prone sites with an odd ratio more than five are V__G, G_G, G_V.
- For the motif, nMn, the tripeptides AGG and GGV present only in myeloid cancer prone sites and not in neutral sites.
- On the other hand, several motifs are preferred at neutral site is G_P (n_M) with an odds ratio of 0.18. In tripeptides, the motifs (nMn) RDI,GQE,SRL,GRH,GDY and GCP are preferred only in neutral sites.
- We have identified the proteins with less number of experimentally known cancer-prone and neutral sites.
- The topological locations of cancer-prone and neutral sites were examined. We observed that 1851 sites are dominated outside of the membrane, whereas 172 residues were found inside of the membrane and 33 residues were located in the membrane region.

V. Web application development

- A web server (MLCanPred) was developed to predict disease prone sites of AML and is freely available at https://ml-canpred.web.app.

CONCLUSION

Overall, in this study by using a series of bioinformatic methods, we identified DEGs, GO, KEGG pathways associated with PTPN11 mutation. Then, we considered a weighted gene co-expression network, and identified the key gene modules and screened out six new biomarkers which are related to AML prognosis and survival. Finally, we classified the disease-causing mutations and developed a web application, which could serve as an effective tool to predict disease prone sites which might be helpful for functional studies and cancer therapeutics.

REFERENCES

Estey E. H. (2020). Acute myeloid leukemia: 2021 update on risk-stratification and management. American journal of hematology, 95(11), 1368–1398. https://doi.org/10.1002/ajh.25975

Siegel R. L., Miller K. D, Fuchs H. E, Jemal A. (2021). Cancer Statistics, 2021. CA: a cancer journal for clinicians, 71(1), 7–33. https://doi.org/10.3322/caac.21654

Vakiti A, Mewawalla P. Acute Myeloid Leukemia. [Updated 2021 Aug 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507875/

Tebbi C. K. (2021). Etiology of Acute Leukemia: A Review. Cancers, 13(9), 2256. https://doi.org/10.3390/cancers13092256

Im, A. P., Sehgal, A. R., Carroll, M. P., Smith, B. D., Tefferi, A., Johnson, D. E., & Boyiadzis, M. (2014). DNMT3A and IDH mutations in acute myeloid leukemia and other myeloid malignancies: associations with prognosis and potential treatment strategies. Leukemia, 28(9), 1774–1783. https://doi.org/10.1038/leu.2014.124

Handschuh L. (2019). Not Only Mutations Matter: Molecular Picture of Acute Myeloid Leukemia Emerging from Transcriptome Studies. Journal of oncology, 2019, 7239206. https://doi.org/10.1155/2019/7239206

Rangaswamy, U., Dharshini, S., Yesudhas, D., & Gromiha, M. M. (2020). VEPAD - Predicting the effect of variants associated with Alzheimer's disease using machine learning. Computers in biology and medicine, 124, 103933. https://doi.org/10.1016/j.compbiomed.2020.103933

Pandey, M., & Gromiha, M. M. (2021). Predicting potential residues associated with lung cancer using deep neural network. Mutation research, 822, 111737. https://doi.org/10.1016/j.mrfmmm.2020.111737

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic acids research, 50(W1), W216–W221. Advance online publication. https://doi.org/10.1093/nar/gkac194

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C., & Chanda, S. K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications, 10(1), 1523. https://doi.org/10.1038/s41467-019-09234-6

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics, 9, 559. https://doi.org/10.1186/1471-2105-9-559

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research, 45(W1), W98–W102. https://doi.org/10.1093/nar/gkx247

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic acids research, 45(W1), W98–W102. https://doi.org/10.1093/nar/gkx247